Presented at the Statistical Society of Canada Annual Meeting, June 1999
To be published in the 1999 Proceedings of the Survey Methods Section of the Statistical Society of Canada

ESTIMATION PROCEDURES FOR THE UNITED STATES NATIONAL RESOURCES INVENTORY

Wayne A. Fuller[1]

ABSTRACT

The United States National Resources Inventory is a survey of land use and land cover conducted on a nationwide sample of approximately 300,000 area segments. Data are collected for the segments as a whole and for points within the segments. Information on acreages in political subdivisions (counties) and in topographic subdivisions (hydrologic units), on water areas, and on federally owned areas is available from external sources. Methods of combining this information and the two levels of sample data to create a user-friendly tabulation file are described.

KEY WORDS: Auxiliary information; Two-phase sample; Small area estimation; Panel survey; Environmental survey.

1. THE UNITED STATES NATIONAL RESOURCES INVENTORY

The *National Resources Inventory* (NRI) is conducted by the United States Natural Resources Conservation Service in cooperation with the Iowa State University Statistical Laboratory. The survey is a panel survey of land use and was conducted in 1982, 1987, 1992, and 1997. In the survey, data are collected on soil characteristics, land use, land cover, wind erosion, water erosion, and conservation practices. The sample is a stratified sample of the nonfederal area of all states and Puerto Rico. The primary sampling units are areas of land called *segments*. The segments vary in size, from 40 acres to 640 acres. Data are collected for the entire segment on items such as urban land and water area. Detailed data on soil properties and land use are collected at a random sample of points within the segment. Generally, there are three points per segment, but 40-acre segments contain two points and the samples in two states contain one point per segment. Some data, such as total land area, federally owned land, and area in large water bodies, are collected on a census basis external to the sample survey. The current sample contains about 300,000 segments and about 844,000 points. See Nusser and Goebel (1997) for a more complete description of the survey. The population of interest for the National Resources Inventory is non-federal land. The NRI database represents the total area of the United States, although very little information is given for points on federal lands.

The sample size is such that direct estimates have acceptable variances for large subdivisions of the surface area. One such subdivision is that for *hydrologic units*. Hydrologic units are drainage areas for major streams. There are about 400 four-digit hydrologic units in the United States. The estimation procedure is designed to reproduce the correct acreage for counties where counties are important political subdivisions. There are about 3,300 counties in the United States. Because the sample must provide consistent acreage estimates for both counties and hydrologic units, the basic tabulation unit is the portion of a four-digit hydrologic unit within a county. This unit is called a *HUCCO*. There are about 5,000 of these units. Some HUCCOs are relatively small and may contain only two segments. Some are relatively large and contain more than 100 segments.

2. ESTIMATION FOR THE NATIONAL RESOURCES INVENTORY

Two possible ways of classifying the surface of the earth are land cover and land use. Land cover is the kind of vegetation, constructed material (such as roads or buildings), or natural material (such as sand, water, or ice)

[1] Wayne A. Fuller, Iowa State University, 218 Snedecor Hall, Ames, Iowa, USA, 50011-1210, waf@iastate.edu.

that actually covers the earth's surface. Categories for land use include categories such as crop production, residential, and wildlife habitat.

In the NRI, all land is placed into mutually exclusive and exhaustive categories, called *coveruse* categories. As the name suggests, the classification is based on both the land cover and the land use. For example, land is classified as urban if it has a certain density of buildings, even if the predominant cover is trees. Roads that are in the rural areas are classified as roads, while roads within the urban area are classified as urban. Other coveruse categories include categories such as cultivated cropland, forest, rangeland, and pastureland.

Analysis variables are created from the data collected at the points. In addition to coveruse, data on crops, cropping practices, slope, erosion, soil properties, land cover, and wildlife habitat are collected at the points.

The sample can be considered to be a two-phase sample with the segment data corresponding to first phase data and the point data corresponding to second phase data. The common estimation procedure for two-phase samples is to assign weights to the second phase units and use the second phase units as the tabulation units. An alternative procedure is to use the first phase units as the tabulation units and to impute the second phase data for the first phase units with no second phase data. Because estimates for relatively small geographic areas are of interest, the imputation approach is used in the NRI.

Estimation for the NRI combines information from a number of sources. The sample data are the segment data and the point data. The segment data elements important for estimation are:

1. acres in large urban; urban areas whose total area is greater than ten acres,
2. acres in small build-up areas; urban areas less than ten acres in total size,
3. acres in small streams; streams and rivers that are less than $1/8^{th}$ of a mile wide,
4. large streams; streams that are greater than $1/8^{th}$ of a mile wide,
5. acres in small water bodies; water bodies less than 40 acres in size,
6. large water bodies; bodies greater than 40 acres in size,
7. acres in farmsteads, and
8. acres in rural roads.

The acres in roads for the two years, 1992 and 1997, were collected for the segments in 1997. Prior to that, the segment data did not contain data on roads. In all cases, the area recorded is the area of that category in the segment.

The second set of major inputs into the estimation process came from a large Geographic Information System (GIS). The GIS information is basically a digital map of the United States. The map contains the boundaries of states and counties, the boundaries of hydrologic units, boundaries of large water areas, where large water areas are water bodies larger than 40 acres and streams greater than one eighth of a mile wide. The system also contains a digital map of federal lands. The GIS information is available for the years 1992 and 1997.

In addition to the digital map, administrative data on acres in the Conservation Reserve Program (CRP) for the years 1992 and 1997 are used as controls in the estimation process. The CRP is a federal program under which land is removed from production for an extended period.

For small area estimation of acres in roads, additional data from the GIS are used. Based on the GIS data for the road network, a variable correlated with acres in roads in rural areas is constructed. A third data input into estimation is population data. The 1992 and 1997 county populations from the United States Census are used to construct variables for small area estimation of change in urban acres.

The estimation process is complicated by the fact that the 1997 data gatherers were permitted to change data for 1982, 1987, and 1992. For example, if the 1997 data gatherer determines that a mistake was made in one of the years, then they are permitted to change earlier data. In fact, they are encouraged to change the data in order to make it internally consistent. Data for earlier years are also changed because data edits improve over time and

edits are expanded when additional data for the year 1997 are collected. Hence, data for 1992 may be changed because of an edit that would not have been used in 1992.

The NRI estimation procedure was developed under several goals. The primary goal was to create a final data set that contains all of the information from the various sources, to the degree that it is operationally possible. It was desired that the final data set be user-friendly in that it is an easy-to-tabulate data set, and it should not be necessary for the user to understand the complexities of the estimation in order to construct estimates. The data set should produce estimates that agree with known data. For example, the weighted sum of data elements in a county should equal the county acres as given by GIS and the sum of data elements in a hydrologic unit should be the acres in that hydrologic unit. In addition, the data set should provide estimates of federal acres, of large bodies, of large streams, and of acres in the CRP that are consistent with the control data.

Although 1997 data elements for the year 1992 need not be the same as data elements recorded in 1992, data users appreciate very much a data set in which the estimates for 1992 from the 1997 data set are very similar to estimates for 1992 from the 1992 data set. Equality is not possible for all variables, but it is highly desirable that the estimates be as similar as possible and be equal for some of the more important data elements. An example of an important data element is the acres of cultivated cropland.

It is desirable that the data set produce reasonable estimates for small areas. The sample size is relatively small for counties, generally on the order of one hundred segments. This means that estimates of rare items, such as acres of urban land or of roads, may have a very large variance at this level. Hence, direct estimates can be relatively unreasonable when compared to known data for the small areas. Thus, small area estimation procedures were adopted for some data elements.

Given the goals, it is clear that the estimation procedure cannot be simple in any absolute sense. But whenever there was a choice, relatively simple procedures were chosen. Furthermore, an estimation procedure with minimal human intervention in the actual estimation process was designed. Clearly, the final estimates were reviewed and anomalies investigated. Generally speaking, the objective was a procedure where anomalies are the result of data inconsistencies, not of choices within the procedure.

Finally, a variance estimation was part of the estimation procedure. As might be inferred from the position of variance estimation in the set of goals, the relatively complicated design, the multiple sources of information used estimation, and limited resources, the variance estimates are based on a number of approximations.

The basic estimation unit, the HUCCO, has been defined as the intersection of counties and hydrologic units. If the estimation procedure is such that the weighted elements in each HUCCO add to the acres in the HUCCO, then the weighted sums produce the correct acres for both counties and hydrologic units. Controls at the HUCCO level are imposed on the total area, on the area in federal lands, and on the area in large water.

Other controls are imposed at the state level. For example, the CRP acres are sometimes very small in some counties and there are not observations on conservation reserve in every county that has CRP. Therefore, a procedure is used that makes the CRP acres close for each county, but the total control is imposed only at the state level. A control is imposed on cropland, pastureland, forest, other rural land, farmsteads, small water, and urban at the state level for 1982, 1987, and 1997, where the control acres are approximately, the acres as reported in the 1992 data set.

To create a data set that contains the information in the GIS data set and the information in the survey sample data set, it is necessary to place the information from the GIS data set into a format compatible with the NRI sample data. To accomplish this, data elements, called *points*, are created where there is a point for the portion of every GIS water body in a HUCCO and a point for a portion of every GIS federal tract in a HUCCO. Thus, if a water body is completely within a HUCCO, one point is created. If the water body crosses the boundary of the county or the boundary of the hydrologic unit, then there is a point for the portion in each HUCCO. The acreage for the water body comes from GIS and other data elements are obtained from the sample data.

To obtain sample data for the GIS water and federal tracts, the survey data are searched. If there is a data point in the survey data that falls in the tract, or if there is more than one point that falls in the tract, then the data from the survey points are assigned to that water body. If there are no survey data for the tract, then the data are imputed for the tract using a point that falls on a similar water body. This imputation procedure is used for points that are in water in 1992 or 1997. Data on changes in water and federal for years prior to 1992 are the data in the 1992 data set.

By far, the largest and most complex operation in estimation is the conversion of segment data into point data. Figure 1 contains a sketch of a 160-acre segment. This is the typical size for a segment in the Midwest. The segment contains 32 acres that were urban in all four years. There are 18 acres that were converted from some other use to urban in 1997. There is also a farmstead in the segment. There is a road along the east side of the PSU, and a road runs through the middle of the segment.

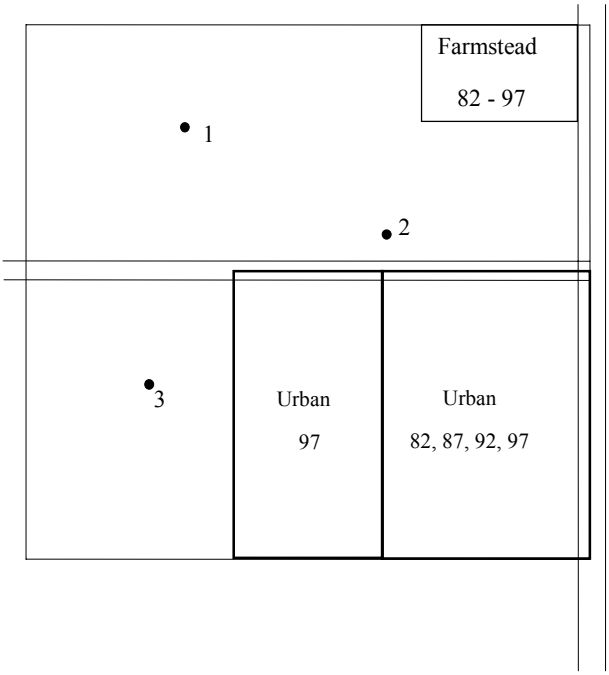

Figure 1. Sketch of an illustrative segment.

The coveruse information available at the estimation stage is given in Table 1 and Table 2. There are 32 acres, in urban for the first three years and 50 acres in 1997. Road information was not collected at the segment level in 1982 and 1987. The road acres are 2.9 acres and 2.7 acres for 1992 and 1997, respectively. Ten acres in farmstead is reported for every year. The total segment acreage is 160. The survey point data are given in Table 2. Point one was pasture all four years. Points two and three are cropland in a two-year rotation of corn and soybeans. The points, although well scattered, do not fall on any of the farmstead or urban land. The acre information by year is the total information that is available at the estimation stage. The segment information is converted into point information by constructing points to represent the four years of data.

A point that is a farmstead every year represents the ten acres in farmstead all four years. Likewise, a point is created for the 32 acres that are urban every year.

Table 1. Acres of segment coveruse by year.

	Coveruse			
	Urban	Roads	Farmsteads	Total Acres
1982	32	NR	10	160
1987	32	NR	10	160
1992	32	2.9	10	160
1997	50	2.7	10	160

Table 2. Coveruse of real points.

	Coveruse		
	Point 1	Point 2	Point 3
1982	Pasture	Corn	Soybeans
1987	Pasture	Soybeans	Corn
1992	Pasture	Corn	Soybeans
1997	Pasture	Soybeans	Corn

Table 3. Coveruse and acres of imputed points representing segment data.

	Imputed Point				
	No. 1	No. 2	No. 3	No. 4	No. 5
1982	Farmstead	Urban	Soybeans	Roads	Roads
1987	Farmstead	Urban	Corn	Roads	Roads
1992	Farmstead	Urban	Soybeans	Roads	Roads
1997	Farmstead	Urban	Urban	Roads	Urban
Acres	10	32	17.8	2.7	0.2

Table 4. Coveruse and acres of real points.

	Real Points		
	Point 1	Point 2	Point 3
1982	Pasture	Corn	Soybeans
1987	Pasture	Soybeans	Corn
1992	Pasture	Corn	Soybeans
1997	Pasture	Soybeans	Corn
Acres	32.433	32.433	32.433

There is no information on the kind of land that was converted to urban use. One of the three real points is chosen at random to represent land converted to urban. In this case, point number three was selected. Hence, the point representing the change has coveruse soybeans, corn, soybeans, and urban in the years 1982, 1987, 1992, and 1997, respectively. A point that is roads every year is created on the basis that there is a minimum of 2.7 acres in roads. A point that goes from roads into urban is also created, because under the NRI definitions, roads in urban areas are classified as urban, not as roads. The four years of segment data are represented by the five points in Table 3. These points are called *imputed points* or *pseudo points*.

The total acres in the segment in farmsteads, urban and roads in 62.7 acres. That leaves 97.3 acres to be allocated to the other uses and one-third of the remainder, 32.43 acres, is assigned to each of the three real points. A total of eight points are used to represent the information that is in the segment and point data for this particular segment.

The estimates of coveruse for each of the years is consistent with the survey information. For example, the shift from non-urban to urban use is a direct estimate, but the previous use of the urban land is an imputed value. Some investigation has been conducted comparing the imputed estimates with the usual two-phase estimates and the imputation procedure seems to work quite well. The number of points available for use as donors in the imputation is often limited. For example, there were two alternatives in the example, cropland and pasture, but in many segments the three points would have been on cropland, and, hence, the only available donors for imputation would have been cropland.

There is another important step in completing the data set. Consider the imputed point that went from cropland to urban. The data available from point number three provides imputed point information for 1982, 1987, and 1992. For data on the urban land in 1997, it is necessary to use a real urban point, a point that is urban in 1997,

as a donor for the urban data for 1997. A hot-deck procedure is used in which a donor unit as similar as possible to the recipient is selected. The rules for similarity include geographical proximity as a criterion. If at all possible, a point within the same HUCCO is chosen. After imputation is complete, eight points represent the data for the segment, and every point has a complete set of information.

After imputation, a point data set exists that contains the segment information and the GIS information. Weights are then assigned to the points. The original weight for a segment is the inverse of the selection probability. In the Midwest, a typical sampling rate is two 160-acre segments selected from a stratum composed of 48 segments. Thus, the original weight for a segment is 24 and the original weight for a point is 1,280 acres. Weights were constructed for the 1992 data set using acre controls for counties and federal land.

In 1997, the weighting procedure began with the point weights from 1992. If the point matched 1992, either a real or imputed point, it received the weight that it had in the 1992 HUCCO data set. Pseudo points with no matching 1992 pseudo points were given the original sampling weight. Real points that had their coveruse for 1992 changed in the 1997 data collection were given a weight equal to one-half of the 1992 weight. The weights for federal and water were determined by the size of the tracts, because the points represent the particular federal and water tracts, as described earlier. The point weights, except the federal and water point weights, are ratio adjusted to the HUCCO acres. The ratio adjusted weights, called *step one weights*, give the correct HUCCO acres and the correct federal and water acres.

The step one weights are used to construct small area estimates for roads and for the change in urban. The model for roads is a components of variance model in which it is assumed that the acres in roads is a function of the GIS measure of road acres. The sampling error can be estimated for most HUCCOs, but because there are a small number of segments in some HUCCOs, the sampling error of the estimated variance is large in those HUCCOs. Therefore, a model was postulated for the sampling variance. It is assumed that the variance depends on the number of segments in the HUCCO and on the true value for the HUCCO.

The estimate of acres in roads is a convex combination of the direct estimate and the model estimate, where the weights are a function of the two variance components. The individual small area estimates are summed and ratio adjusted so that the sum is equal to the direct state estimate. This means that the state estimate is a design unbiased estimate of acres in roads, but the estimates for the local areas use information from the GIS. The weights of the original observations are ratio adjusted to give the individual small-area HUCCO estimates.

A similar small-area model is used for urban land, where the model is for the change in urban land from 1992 to 1997 and the explanatory variables in the model are functions of the 1992 and 1997 Census county population. The vector of regression coefficients is estimated with generalized least squares. The variance components are estimated and the convex combination of model estimate and direct estimate formed in much the same way as for roads.

The acres in the CRP are available from administrative sources by county for 1992 and 1997. The acres in CRP for 1992 and 1997 are used as controls. The acres in counties that have CRP points are ratio adjusted upward to adjust for counties which have no sample observations. Then the weights on CRP points are ratio-adjusted to give the state control numbers.

A raking operation is used to bring the state coveruse estimates for 1982, 1987, and 1992 for the 1997 data set into agreement with the corresponding estimates provided by the 1992 data set. Because the acres in roads, water and federal land available from GIS are not exactly the same as the acres of those coveruses in the 1992 data set, not all coveruses agree exactly with the corresponding acres reported in the 1992 data set. The raking operation alternatively controls on HUCCO acres and on the state acres for the relevant coveruses for the years 1982, 1987, and 1992.

The final step in weight construction is to round all of the weights to hundreds of acres. This means that tables created using rounded weights are internally consistent. Strata and clusters are created for variance estimation. The variance estimation strata are not the original geographic strata, partly for confidentiality reasons and partly to reflect the fact that external information was used in the estimation process.

The final data set is a relatively simple data set. It contains the stratum identification and the cluster identification for variance estimation and a kind-of-point indicator that classifies the point as a real point or a point created to represent segment information. The weight and the data elements complete the data for a point. This data set accomplishes the purpose of summarizing all of the available information into a form that is relatively easy to use.

REFERENCES

Breidt, F. J., McVey, A., and Fuller, W. A. (1996-97). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, 49, 79-90.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

Nusser, S. M. and Goebel, J. J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4, 181-204.